Classification and Topic Modeling tasks on textual datasets

Alberto Arienti, Sofia Turroni
University of Milan Bicocca

Abstract
In this project, the analysis of a textual dataset is addressed using text mining techniques. The main objective is to preprocess the textual data, classify the content into predefined categories, and extract the main themes (topic modeling) to obtain a structured representation of the information. Preprocessing includes cleaning, tokenization, and removal of irrelevant elements. Classification is implemented using machine learning algorithms, while the identification of main topics exploits topic modeling methods such as Latent Dirichlet Allocation (LDA). This approach makes it possible to transform unstructured data into structured, meaningful information.

Keywords: Text Mining, Topic Modeling, Text classification, Machine Learning

Contents

1 Introduction, dataset description & objectives

Textual data analysis is gaining increasing relevance in various fields due to its ability to extract useful and meaningful information from large volumes of unstructured text. In this project, we focus on the application of text mining techniques, in particular **text classification** and **topic modeling**, to analyze and interpret a text dataset.

1.1 Dataset description

The dataset we decided to use in this project is called **PubMed 200k RCT**, was selected by *Kaggle*, and is available at the following link: PubMed. PubMed 200k RCT is a dataset based on PubMed for sequential sentence classification. The dataset consists of approximately 200,000 abstracts of randomized con-

trolled trials.
Each of these abstracts is divided several times according to the number of sentences from which it is composed and each abstract sentence is labelled with its role in the abstract using one of the following classes: BACKGROUND, OBJECTIVE, METHOD, RESULT, or CONCLUSION. Eanch abstract may have several labels of the same type e.g., the same abstract may have two or more instances labelled as RESULT.
From the link there are several directories containing different files, for our project we chose the file `Train.csv` contained in the directory `PubMed_20k_RCT`, which instead of 200,000 abstracts contains 20,000, a sufficient number for our purposes.
The file consists of 180,040 lines, where each line identifies a part of medical publication abstracts. While the columns in the file are:

- `abstract_id`: uniquely identifies which abstract the text fragment comes from;
- `line_id`: identifies the line of the abstract, is composed of the variable *abstract_id* and the line number of the reference abstract;
- `abstract_text`: text part of the abstract;
- `line_number`: line number of the abstract;
- `total_lines`: total number of lines into which the abstract was divided.
- `target`: identifies what the text contained in the *abstract_text* variable refers to, it may take five classes: BACKGROUND, OBJECTIVE, METHODS, RESULTS and CONCLUSIONS.

1.2 Objectives

The main objective is twofold:

1. To pre-process and classify the different instances of the text according to specific categories.
2. Extract the main themes (topics) from the documents through thematic modeling techniques.

For the text classification task, we propose to develop a model capable of distinguishing various instances of text according to their labels. This will make it possible to automate the categorization of text and assess its accuracy.
In the topic modeling task, document parts will be aggregated to form a larger corpus. From this corpus, we will use techniques such as Latent Dirichlet Allocation (LDA) to identify the main topics discussed in the documents. This analysis will allow us to extract semantic patterns and understand the main topics in the dataset.
In summary, the project aims to combine machine learning and thematic modeling methods to exploit the full potential of textual data, offering insights into both classification and the discovery of hidden topics.

---

2 Text Classification

As described above, the dataset consists of scientific papers' abstracts divided into several sentences, each of which is labeled in the `target` variable. The objective of classification is therefore to automatically identify the correct category for each text instance, using supervised machine learning techniques. Once the text has been pre-processed and vectorized, we will train different classifiers and then choose the one that performs best.
In a real-world application, it might be useful if, in case one wanted to know the methodology rather than the results or conclusions of a document, these could be highlighted by the model without reading the entire document.

2.1 Text pre-processing

The pre-processing of the text is a fundamental step that includes cleaning and transforming the text into a format suitable for analysis; these operations will be performed on the `abstract_text` column of the dataset.
Here is a list of the different pre-processing steps:

1. Transformation of words: the first pre-processing step we applied is to transform the word *p-value* into *p_value*, since after an exploratory phase we noticed that the word appeared very frequently, this will be useful in the future to make the word appear as a unique token.

2. Removal of punctuation and numbers: punctuation and numbers, not contributing directly to the semantic meaning of the text for our purposes, were removed.

3. Removal of stop words: for this step, we used the default list of stop words provided by the **NLTK (Natural Language Toolkit)** library. Stop words (common words such as *'the'*, *'and'*, *'but'*, *'of'*), which tend to appear frequently without contributing significant semantic information, were removed.

4. Tokenization, the text was then broken down into tokens, i.e. individual linguistic units (typically words) in order to treat the text as a sequence of discrete elements.

5. Lemmatization: this is the last step , the process of lemmatization was applied, which reduces each word to its basic form (lemma). To ensure the best possible performance, we used a POS tagging method to previously tell the lemmatizer which type of word is he dealing with, providing it the information for the correct lemmatization.

A column called `tokens` will be added to the data frame, which will contain the tokens calculated from these different phases. This cleaning process reduced the dimensionality of the text, eliminating redundant information and retaining only the most relevant elements, for a single instance this will then consist of the text contained in the variable `abstract_text` but without stop words, punctuation and numbers and lemmatized.

2.2 Vectorization

In order to make the text data readable by the algorithms, we chose to use the **TF-IDF (Term Frequency-Inverse Document Frequency)** vectorization technique. The decision to use this type of vectorization is due to the fact that it provides an effective representation, balancing the frequency and relevance of terms, is computationally simpler than other vectorization techniques while maintaining a high quality of representation, and at the same time is more robust than others such as the Count Vectorizer. TF-IDF in the context of scientific papers could be the best option to better highligh specific and rare terms.
In the implementation of TF-IDF, we configured the following parameters:

- Lowercase (lowercase=True): to ensure uniformity in the text, all words were converted to lower case before being vectorized. This step prevents terms written in lower and upper case from being treated as different.

- Max Document Frequency (max_df=0.75): terms present in more than 75% of the documents were excluded, as they were considered too frequent and therefore not very informative.

- Minimum Document Frequency (min_df=50): terms present in less than 50 documents were excluded, as considered to be too rare to be informative for the classification task.

2.3 Models training and evaluation

2.3.1 Training

To tackle the text classification task, several machine learning models have been tested in order to identify the one with the best performance. The classifiers considered were chosen for their wide application in classification problems. They are listed below with a brief description of each algorithm:

- **Decision Tree**, are simple, interpretable models that subdivide the dataset according to specific criteria;
- **Random Forest**, is an ensemble method based on decision trees;
- **Logistic Regression**, is a linear classifier that uses a logistic function to model the probability of belonging to a class;
- **Support Vector Classifier**, a variant of the Support Vector Machine (SVM) that tries to find an optimal hyperplane to separate data into different classes;
- **Multinomial Naive Bayes**, a variant of the Naive Bayes model.

Before testing the different classifiers, the dataset was divided into two subsets:

- Training set (80%): used for training the models.
- Test set (20%): used for evaluating the performance of the models on unseen data.

Each model was trained using the training set and, in order to evaluate the performance of the models on the test set, several standardized metrics were used for classification. For computational reasons, the support vector classifier was evaluated with approximately half of the observations (70,000 observations for the training set and 10,000 for the test set).

2.3.2 Evaluation

To evaluate the performance of the developed models, we used several metrics commonly adopted in machine learning and classification. Specifically, the main metrics used are:

- **Accuracy**: measures the percentage of correctly classified instances out of the total. It provides an overview of the model's performance;
- **Recall**: measures the proportion of true positives to the total number of true positive samples;
- **Precision**: defines the proportion of true positives to the total number of positive predictions;
- **F1-Score**: represents the harmonic mean between precision and recall.

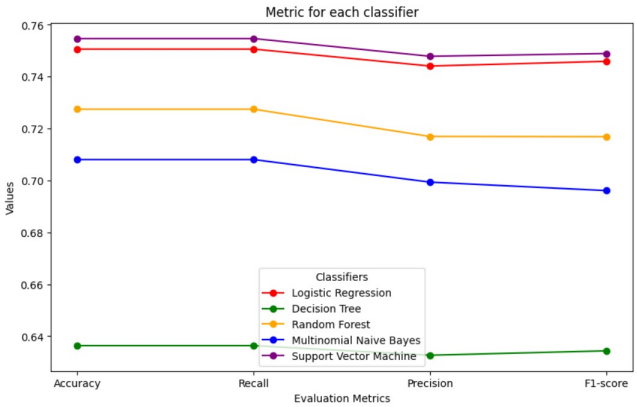

Figure 1: Comparison of key evaluation metrics between the chosen models.

The Figure 1 shows a line plot representing the values of the evaluation metrics chosen for each trained classifier. As it shows, the Support Vector Machine and Logistic Regression stood out as the best classifiers, showing high and consistent results for all metrics, especially the former taking into account that it also had training limitations compared to the other classifiers. In contrast, Decision Tree performed the worst, suggesting that it might not be suitable for this task.
Our choice will therefore fall back on the use of Support Vector Classifier or Logistic Regression, for these two we also extracted the heatmap of the confusion matrix, Figure 2 and 3.

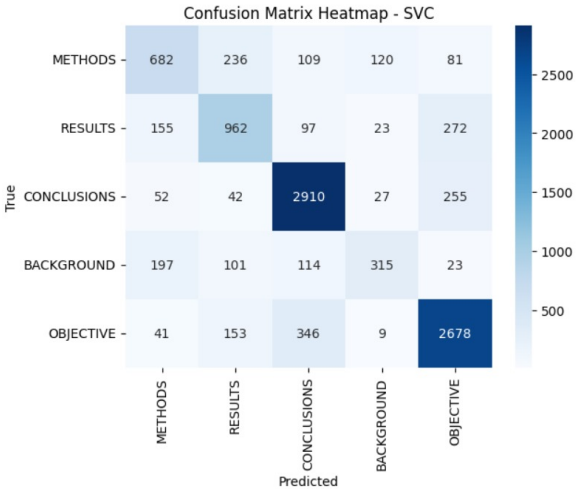

Figure 2: Confusion matrix for Support Vector Classifier.

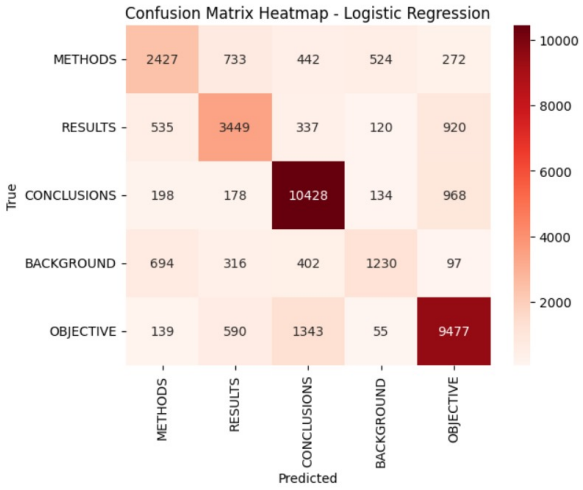

Figure 3: Confusion matrix for Logistic Regression.

Heatmaps in Figures 2, 3 shows, for each instance and its label, the correct label present in the test set and the predicted class by the model. Both models showed solid performance, with a high concentration of values along the main diagonal, which represents the correct predictions. From the heatmap we see that both models tend to predict the class OBJECTIVE as CONCLUSION and the class RESULT or CONCLUSION as OBJECTIVE. Although the Support Vector Classifier has better metrics our choice was to use **Logistic Regression** as the classifier for our project. Our choice is justified by the fact that it is faster and less computationally demanding, the classification choices are more explainable and the results are not significantly inferior but very similar.
The complete confusion matrix for the chosen classifier is shown in the table 1. By analyzing it we can notice that the classes METHODS and RESULTS are the ones which the model recognize with the highest precision and recall, meaning that it can correctly identify the most of them, while BACKGROUND and OBJECTIVE show a very poor value in terms of Recall. This is probably due to the class imbalance, where the first two classes alone represent almost the 65% of the total instances, while BACKGROUND and OBJECTIVE are respectively 12.1% and 7.7%.

	Precision	Recall	F1-Score	Support
BACKGROUND	0.61	0.55	0.58	4398
CONCLUSIONS	0.65	0.64	0.65	5361
METHODS	0.81	0.88	0.84	11996
OBJECTIVE	0.60	0.45	0.51	2739
RESULTS	0.81	0.82	0.81	11604
Accuracy		0.75		36008
Macro avg	0.69	0.67	0.68	36008
Weighted avg	0.74	0.75	0.75	36008

Table 1: Classification Report

3 Topic modeling

In this section, the main topics within the corpus of abstracts will be identified, and papers with similar topics will be grouped together. The algorithms used to address this challenge are Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA) and Hierarchical Dirichlet Process (HDP).
The application of topic modeling techniques in this specific context can be useful in building smarter search systems, by associating documents with main topics, one can improve search and find relevant documents even if they do not contain the same keywords.

3.1 Dataset modification

Before applying algorithms to extract topics from the tests it was necessary to apply a modification to the dataset. As previously illustrated, each abstract is subdivided into a certain number of sentences identified by the variable `total_lines`, we concatenated the text by grouping it by each `abstract_id` and sorted by `line_number`. In this way we went from a total of 180040 lines to 15000, where each instance now represents the complete corpus of each abstract. The same operation was carried out for the `tokens` column, having so all the tokens of the complete abstract.
Now the data frame we are going to use will therefore only consist of three columns: `abstract_id`, `abstract_text` and `tokens`, the only ones relevant to the analysis in question.

3.2 Application of topic modeling

For the application of topic modeling, the first essential step is the *creation of a dictionary* to represent the relevant keywords in the documents. The process for constructing the dictionary consisted of the following steps:

1. From the available tokens, only words were selected that:
   - Appear in at least five documents: to remove too rare or specific terms that do not contribute to the formation of coherent topics.
   - Appear in less than 75% of the documents: to exclude too common and generic words that do not help differentiate topics.
2. Selection of the most relevant keywords: after applying filters, the 100,000 best words were selected according to their relevance.

The creation of a filtered dictionary was a crucial step for the effectiveness of topic modeling; it allows us to reduce noise, focus on the most meaningful words and optimize the model. This dictionary can now be used to represent documents in a format suitable for algorithms, thanks to this preparation, the model will be able to identify the main themes clearly and consistently.
For topic modeling, we applied two main techniques: **Latent Semantic Analysis (LSA)** and **Latent Dirichlet Allocation (LDA)**. In order to choose the optimal number of topics, we used 30% of the dataset, chosen as a representative sample to ensure efficient computational management without compromising the variety of topics. The LDA model was evaluated against perplexity, a metric that measures how well the model is able to represent the dataset, and coherence, a measure of the semantic coherence of the main words in each topic. The LSA model was evaluated with coherence as it reduces the dimensionality of the data (e.g. words and documents) to capture latent relationships, but not dealing directly with the probability of word sequences, it is not possible to calculate a perplexity in the strict sense with LSA.

3.2.1 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a probabilistic model that generates topics as probability distributions over terms. Figures 4 and 5 show the values of perplexity and consistency as a function of the number of topics (5 to 40).
Perplexity rises significantly from topic number 7 to 30, where it then falls again; the optimum number of topics will have to be sought within this range, as the best perplexity is the one which is more closer to 0. Consistency tends to rise gradually along with topics, with a notable surge at topic number 29, where it approaches 0.40 without then rising much. Combining the results from both perplexity and coherence score, the optimal number of topic seems to be 29, as the perplexity reaches one of its peak and the value of coherence is strong. (To visualize topics, see appendix B at the end of the article)

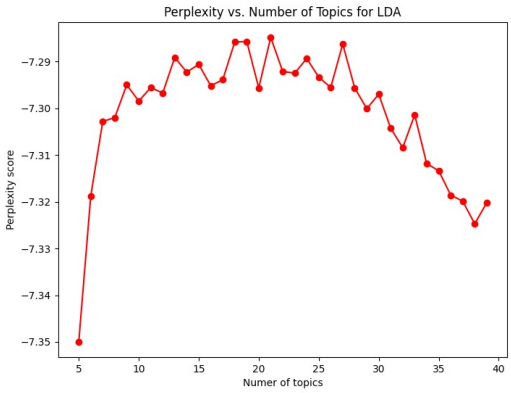

Figure 4: Perplexity in relation to the number of topics for the LDA model.

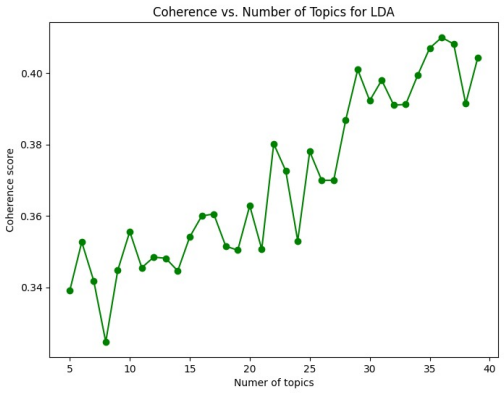

Figure 5: Coherence in relation to the number of topics for the LDA model.

3.2.2 Latent Semantic Analysis

Latent Semantic Analysis (LSA) is based on singular value decomposition (SVD), the model reduces the text size space to identify latent relationships between terms and documents. Since it is not possible to calculate the perplexity for the LSA model, we only evaluated it with coherence where from Figure 6

we can see that it reaches a maximum value of about 0.4 with 7 topics.
As the number of topics where coherence reaches the maximum value is relatively low like other evaluations, we extracted the topics and evaluated them with a human approach. and in the LSA model they are very generic and similar to each other (see appendix A at end of article) despite being much fewer in number than in LDA.

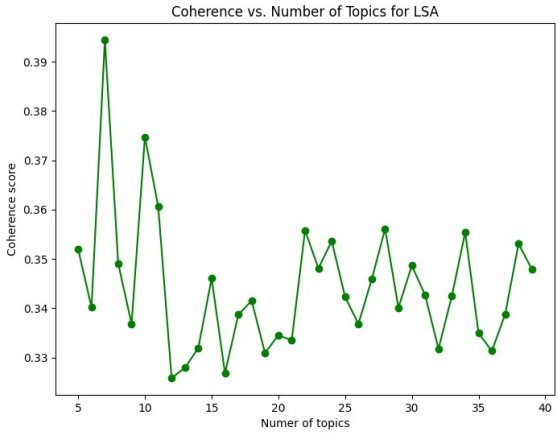

Figure 6: Choerence in relation to the number of topics for the LSA model.

3.3 Choice of model

After exploring various topic modeling techniques, we decided to adopt the **Latent Dirichlet Allocation (LDA) model with 29 topics**, as opposed to the Latent Semantic Analysis (LSA) model which uses only 7 topics, based on a thorough analysis of the results obtained, which also included a human approach.
LDA produced topics that were more distinct and interpretable than those generated by LSA. In particular, the LSA topics were perceived as very generic and unclear, with much overlap between the various topic groups. The topics generated by LSA were in fact similar to each other, which made it difficult to identify distinct topics or gain useful insights from the model.
LDA produced topics that were more distinct and interpretable than those generated by LSA. In particular, the LSA topics were perceived as very generic and unclear, with much overlap between the various topic groups. The topics generated by LSA were in fact similar to each other, which made it difficult to identify distinct topics or to obtain useful insights from the model.
In conclusion, the choice of LDA with 29 topics was motivated by its ability to generate clearer, more distinct and useful topics. The optimal number of topics can vary depending on the specific task; however, in this case, the choice of 29, while relatively large, allows for meaningful interpretation. Appendix B provides the top 10 words for each topic, which facilitated an initial analysis revealing three major thematic groups:

- **Patient Care and Interventions**: Topics 5, 6, 7, 11, 16, 17, 18, 21, 24, 27, 25.
- **Physical and Medical Conditions**: Topics 0, 3, 4, 8, 12, 14, 20, 28.
- **Testing and Outcomes**: Topics 1, 2, 9, 10, 13, 15, 19, 22, 23, 26

An alternative approach might involve reduc-

ing the number of topics; however, given the specialized nature of the dataset, which comprises medical scientific papers sharing certain domain-specific characteristics, retaining a relatively high number of topics ensures more granular and specific topic categorization. This approach minimizes the manual effort required for interpretation and enhances the clarity of thematic differentiation across papers.
To further assess the quality of the LDA results, we examined the distribution of topics across the dataset, as shown in Figure 7. The x-axis represents the 29 identified topics, while the y-axis indicates their proportional occurrence within the dataset. The results demonstrate a well-balanced topic distribution, with no extreme disparities in frequency. While certain topics, such as Topic 25, Topic 21, and Topic 6, appear more frequently in abstracts, the differences are not substantial enough to indicate suboptimal topic modeling performance.

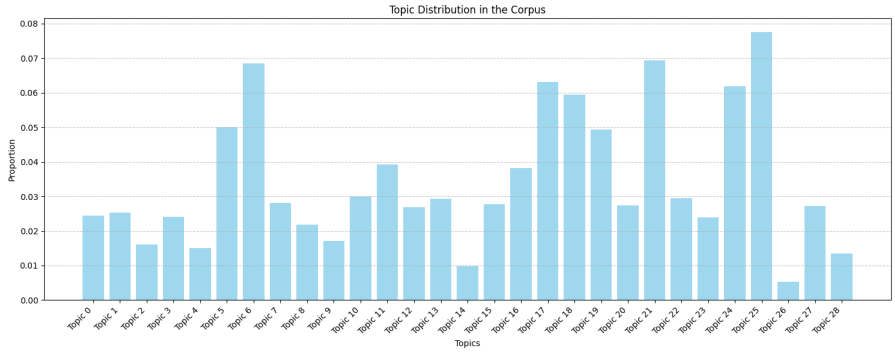

Figure 7: Choerence in relation to the number of topics for the LSA model.

4 Conclusions

The aim of the project was to apply two text mining tasks: classification and topic modeling. Below, we present the conclusions for each task.
For the classification task, we explored various classifiers, evaluating their performance based on their ability to classify parts of abstracts according to their class. After training and testing various models, we chose Logistic Regression as the best classifier for this specific application. This model showed good generalization capability, with solid accuracy in class prediction. The classifier made it possible to identify sentence categories within abstracts effectively.
After classifying the sentences of the abstracts, we combined the sentences belonging to the same class to form the complete abstracts. Next, we applied the Latent Dirichlet Allocation (LDA) model to extract topics from the resulting documents, using a number of 29 topics. LDA provided a good semantic representation of the data, with distinct and interpretable topics reflecting the predominant themes within the abstracts. The choice of 29 topics was motivated by the need to achieve greater granularity than more generic solutions, such as those obtained through dimensionality reduction techniques.
In summary, the project achieved good results in both tasks: the Logistic Regression classifier effectively divided the papers into relevant classes, while topic analysis using LDA offered a clear and detailed view of the topics covered in the abstracts. Both approaches proved suitable for the objectives set, allowing efficient data management and in-depth content analysis.

Appendix A

Topic 0: - group (Weight: 0.5358) - patient (Weight: 0.3958) - treatment (Weight: 0.2021) - P (Weight: 0.1891) - study (Weight: 0.1756) - control (Weight: 0.1592) - use (Weight: 0.1188) - compare (Weight: 0.1132) - week (Weight: 0.1102) - month (Weight: 0.1092)

Topic 1: - group (Weight: -0.7438) - patient (Weight: 0.4571) - study (Weight: 0.1425) - trial (Weight: 0.1170) - placebo (Weight: 0.0876) - year (Weight: 0.0872) - P (Weight: -0.0837) - outcome (Weight: 0.0815) - CI (Weight: 0.0795) - risk (Weight: 0.0789)

Topic 2: - patient (Weight: 0.6097) - intervention (Weight: -0.3663) - control (Weight: -0.1918) - use (Weight: -0.1629) - group (Weight: 0.1560) - study (Weight: -0.1450) - participant (Weight: -0.1252) - effect (Weight: -0.1216) - child (Weight: -0.1168) - trial (Weight: -0.1093)

Topic 3: - P (Weight: 0.4466) - treatment (Weight: 0.4290) - intervention (Weight: -0.3116) - patient (Weight: -0.3029) - placebo (Weight: 0.2540) - week (Weight: 0.2225) - group (Weight: -0.1778) - mg (Weight: 0.1678) - care (Weight: -0.1350) - p (Weight: -0.1323)

Topic 4: - P (Weight: -0.6479) - treatment (Weight: 0.5318) - week (Weight: 0.2098) - p (Weight: 0.1446) - v (Weight: -0.1420) - risk (Weight: -0.0991) - placebo (Weight: 0.0967) - high (Weight: -0.0958) - mg (Weight: 0.0928) - patient (Weight: -0.0826)

Topic 5: - treatment (Weight: 0.4700) - p (Weight: -0.3547) - P (Weight: 0.3134) - placebo (Weight: -0.2516) - mg (Weight: -0.2331) - study (Weight: -0.2153) - dose (Weight: -0.2020) - intervention (Weight: 0.1929) - month (Weight: 0.1442) - care (Weight: 0.1346)

Topic 6: - CI (Weight: 0.3145) - month (Weight: 0.3034) - pain (Weight: -0.2816) - year (Weight: 0.2776) - v (Weight: 0.1987) - patient (Weight: -0.1833) - study (Weight: -0.1750) - risk (Weight: 0.1660) - use (Weight: -0.1651) - p (Weight: 0.1530)

Appendix B

Topic 0: - exercise (Weight: 0.0344) - group (Weight: 0.0250) - muscle (Weight: 0.0200) - P (Weight: 0.0199) - bone (Weight: 0.0181) - increase (Weight: 0.0149) - asthma (Weight: 0.0132) - training (Weight: 0.0131) - oxygen (Weight: 0.0126) - pressure (Weight: 0.0124)

Topic 1: - smoking (Weight: 0.0174) - use (Weight: 0.0169) - test (Weight: 0.0139) - participant (Weight: 0.0116) - report (Weight: 0.0105) - woman (Weight: 0.0102) - smoker (Weight: 0.0094) - effect (Weight: 0.0088) - study (Weight: 0.0088) - memory (Weight: 0.0086)

Topic 2: - mm (Weight: 0.0460) - eye (Weight: 0.0403) - P (Weight: 0.0241) - visual (Weight: 0.0219) - month (Weight: 0.0205) - mean (Weight: 0.0190) - laser (Weight: 0.0161) - implant (Weight: 0.0154) - group (Weight: 0.0150) - PD (Weight: 0.0131)

Topic 3: - glucose (Weight: 0.0334) - insulin (Weight: 0.0311) - diabetes (Weight: 0.0306) - IL (Weight: 0.0207) - type (Weight: 0.0177) - P (Weight: 0.0162) - control (Weight: 0.0141) - patient (Weight: 0.0140) - level (Weight: 0.0130) - fast (Weight: 0.0107)

Topic 4: - infection (Weight: 0.0460) - antibiotic (Weight: 0.0178) - skin (Weight: 0.0165) - rate (Weight: 0.0158) - lens (Weight: 0.0134) - use (Weight: 0.0112) - HCV (Weight: 0.0099) - G (Weight: 0.0089) - hepatitis (Weight: 0.0085) - donor (Weight: 0.0083)

Topic 5: - treatment (Weight: 0.0366) - score (Weight: 0.0310) - patient (Weight: 0.0249) - symptom (Weight: 0.0214) - week (Weight: 0.0207) - improvement (Weight: 0.0135) - P (Weight: 0.0131) - placebo (Weight: 0.0123) - baseline (Weight: 0.0113) - group (Weight: 0.0111)

Topic 6: - patient (Weight: 0.0508) - CI (Weight: 0.0160) - trial (Weight: 0.0154) - year (Weight: 0.0145) - v (Weight: 0.0119) - primary (Weight: 0.0111) - group (Weight: 0.0108) - event (Weight: 0.0096) - p (Weight: 0.0093) - follow (Weight: 0.0092)

Topic 7: - patient (Weight: 0.0351) - survival (Weight: 0.0291) - month (Weight: 0.0191) - chemotherapy (Weight: 0.0174) - progression (Weight: 0.0137) - free (Weight: 0.0127) - overall (Weight: 0.0126) - arm (Weight: 0.0124) - P (Weight: 0.0123) - HR (Weight: 0.0116)

Topic 8: - child (Weight: 0.0867) - age (Weight: 0.0279) - infant (Weight: 0.0266) - vaccine (Weight: 0.0232) - parent (Weight: 0.0179) - year (Weight: 0.0173) - month (Weight: 0.0153) - vaccination (Weight: 0.0097) - receive (Weight: 0.0095) - HIV (Weight: 0.0091)

Topic 9: - acid (Weight: 0.0264) - mg (Weight: 0.0252) - concentration (Weight: 0.0236) - study (Weight: 0.0203) - h (Weight: 0.0165) - day (Weight: 0.0157) - exposure (Weight: 0.0151) - plasma (Weight: 0.0150) - AUC (Weight: 0.0149) - subject (Weight: 0.0137)

Topic 10: - dose (Weight: 0.0510) - day (Weight: 0.0242) - placebo (Weight: 0.0219) - study (Weight: 0.0199) - dos (Weight: 0.0150) - g (Weight: 0.0138) - injection (Weight: 0.0133) - safety (Weight: 0.0132) - receive (Weight: 0.0131) - mg (Weight: 0.0120)

Topic 11: - group (Weight: 0.1445) - treatment (Weight: 0.0489) - P (Weight: 0.0359) - control (Weight: 0.0262) - two (Weight: 0.0200) - acupuncture (Weight: 0.0176) - day (Weight: 0.0142) - patient (Weight: 0.0141) - case (Weight: 0.0138) - score (Weight: 0.0125)

Topic 12: - patient (Weight: 0.0430) - heart (Weight: 0.0210) - P (Weight: 0.0172) - cardiac (Weight: 0.0147) - renal (Weight: 0.0129) - artery (Weight: 0.0125) - function (Weight: 0.0119) - coronary (Weight: 0.0108) - failure (Weight: 0.0097) - cardiovascular (Weight: 0.0097)

Topic 13: - risk (Weight: 0.0428) - associate (Weight: 0.0273) - age (Weight: 0.0259) - year (Weight: 0.0257) - CI (Weight: 0.0245) - factor (Weight: 0.0244) - model (Weight: 0.0213) - analysis (Weight: 0.0177) - use (Weight: 0.0157) - baseline (Weight: 0.0153)

Topic 14: - sleep (Weight: 0.0491) - alcohol (Weight: 0.0449) - drinking (Weight: 0.0219) - CPAP (Weight: 0.0160) - IOL (Weight: 0.0157) - R (Weight: 0.0130) - apnea (Weight: 0.0109) - use (Weight: 0.0106) - transplant (Weight: 0.0099) - PM (Weight: 0.0097)

Topic 15: - woman (Weight: 0.0289) - day (Weight: 0.0260) - outcome (Weight: 0.0251) - group (Weight: 0.0208) - hospital (Weight: 0.0156) - hour (Weight: 0.0149) - stroke (Weight: 0.0144) - pregnancy (Weight: 0.0134) - blood (Weight: 0.0114) - delivery (Weight: 0.0103)

Topic 16: - patient (Weight: 0.0201) - use (Weight: 0.0192) - group (Weight: 0.0191) - test (Weight: 0.0148) - study (Weight: 0.0124) - training (Weight: 0.0090) - score (Weight: 0.0083) - student (Weight: 0.0082) - p (Weight: 0.0071) - time (Weight: 0.0067)

Topic 17: - group (Weight: 0.0484) - patient (Weight: 0.0284) - surgery (Weight: 0.0197) - postoperative (Weight: 0.0149) - P (Weight: 0.0138) - pain (Weight: 0.0134) - time (Weight: 0.0097) - use (Weight: 0.0094) - significantly (Weight: 0.0094) - Group (Weight: 0.0094)

Topic 18: - group (Weight: 0.0379) - patient (Weight: 0.0300) - pain (Weight: 0.0179) - surgery (Weight: 0.0105) - study (Weight: 0.0102) - month (Weight: 0.0101) - compare (Weight: 0.0100) - difference (Weight: 0.0097) - follow (Weight: 0.0097) - use (Weight: 0.0091)

Topic 19: - level (Weight: 0.0294) - group (Weight: 0.0148) - serum (Weight: 0.0141) - effect (Weight: 0.0138) - P (Weight: 0.0135) - increase (Weight: 0.0133) - concentration (Weight: 0.0129) - placebo (Weight: 0.0125) - C (Weight: 0.0120) - study (Weight: 0.0120)

Topic 20: - weight (Weight: 0.0410) - body (Weight: 0.0254) - diet (Weight: 0.0214) - intake (Weight: 0.0203) - P (Weight: 0.0199) - loss (Weight: 0.0160) - mass (Weight: 0.0159) - fat (Weight: 0.0156) - BMI (Weight: 0.0138) - food (Weight: 0.0133)

Topic 21: - intervention (Weight: 0.0348) - care (Weight: 0.0248) - trial (Weight: 0.0153) - control (Weight: 0.0139) - health (Weight: 0.0129) - cost (Weight: 0.0116) - month (Weight: 0.0106) - outcome (Weight: 0.0098) - use (Weight: 0.0094) - study (Weight: 0.0090)

Topic 22: - use (Weight: 0.0213) - image (Weight: 0.0183) - patient (Weight: 0.0128) - study (Weight: 0.0117) - CT (Weight: 0.0111) - volume (Weight: 0.0109) - MRI (Weight: 0.0089) - size (Weight: 0.0086) - method (Weight: 0.0074) - diagnostic (Weight: 0.0073)

Topic 23: - effect (Weight: 0.0200) - stimulation (Weight: 0.0187) - motor (Weight: 0.0125) - subject (Weight: 0.0110) - response (Weight: 0.0108) - IOP (Weight: 0.0104) - sham (Weight: 0.0103) - study (Weight: 0.0091) - healthy (Weight: 0.0084) - condition (Weight: 0.0084)

Topic 24: - patient (Weight: 0.0442) - treatment (Weight: 0.0325) - week (Weight: 0.0309) - placebo (Weight: 0.0293) - group (Weight: 0.0254) - mg (Weight: 0.0231) - study (Weight: 0.0151) - receive (Weight: 0.0132) - event (Weight: 0.0126) - adverse (Weight: 0.0108)

Topic 25: - intervention (Weight: 0.0213) - group (Weight: 0.0171) - control (Weight: 0.0151) - physical (Weight: 0.0127) - study (Weight: 0.0106) - effect (Weight: 0.0103) - activity (Weight: 0.0102) - program (Weight: 0.0098) - outcome (Weight: 0.0095) - measure (Weight: 0.0082)

Topic 26: - HF (Weight: 0.0473) - CRT (Weight: 0.0195) - HRQOL (Weight: 0.0168) - pace (Weight: 0.0160) - SA (Weight: 0.0135) - hot (Weight: 0.0131) - SMS (Weight: 0.0131) - ASD

(Weight: 0.0116) - use (Weight: 0.0113) - ICD (Weight: 0.0109)

Topic 27: - group (Weight: 0.0387) - control (Weight: 0.0173) - p (Weight: 0.0152) - significant (Weight: 0.0136) - use (Weight: 0.0128) - month (Weight: 0.0120) - study (Weight: 0.0115) - treatment (Weight: 0.0111) - compare (Weight: 0.0099) - clinical (Weight: 0.0090)

Topic 28: - cancer (Weight: 0.0852) - cell (Weight: 0.0375) - breast (Weight: 0.0364) - prostate (Weight: 0.0256) - CD (Weight: 0.0250) - woman (Weight: 0.0237) - expression (Weight: 0.0180) - biopsy (Weight: 0.0159) - RT (Weight: 0.0154) - positive (Weight: 0.0133)